Processing Data

The difference between data and information

All computers can do is recognise two distinct physical states; essentially they can understand whether a switch is on or off.

The brain of the computer, the *CPU (Central processing unit)* consists of several million tiny electronic switches called ***transistors.***

Data is the term used to describe the ***information*** represented by groups of on/off switches.

Even though the words ***data*** and ***information*** are often used interchangeably, there is an important distinction between them. ***Data*** consists of the raw numbers that computers organise to produce ***information***.

How computers represent data

To a computer everything is a number. Numbers are numbers, letters are numbers, sound and pictures are numbers. Even the computers own instructions are numbers. A string of alphabet characters such as a sentence looks just like a string of ones and zeros to a computer.

People usually use the ***decimal number system (base 10)*** to represent numbers, so called because 10 symbols are available (0 – 9). To represent a number above 9 you must use 2 or more symbols together.

A computer however has only 2 possible states available to represent data – on or off. When a switch is off it is represented by a 0, when it is on it is represented by a 1. Because only 2 values are available the system is said to be ***base 2*** or the ***binary number system.*** When a computer represents data greater than 1 it does so by using 2 or more symbols.

Bits and Bytes

The value represented by each switch state (whether on or off) is called a ***bit.*** The term is a combination of *bi*nary dig*it*. A bit is the smallest possible unit of data that a computer can recognise or use. To represent anything meaningful the computer uses bits in groups.

A group of 8 bits is called a ***byte***. With one byte the computer can represent 256 different symbols or characters because the 8 1s and 0s in a byte can be combined in 256 different ways.

The byte is extremely important because there are enough 8-bit combinations to represent all the characters on a keyboard.

Text Codes

Early programmers realised that they needed a standard ***text code*** to represent the alphabet, punctuation marks and other symbols.

EBCDIC, ASCII and ***Unicode*** are 3 of the most popular text code systems invented.

EBCDIC

The ***BCD (binary coded decimal)*** system, defined by IBM for it's early computers was the first complete system to represent symbols with bits.

Used 6-bit codes → maximum of 64 possible symbols.
Only used uppercase letters and very few other symbols → short lived.

Need to represent more characters → development of ***EBCDIC (extended binary coded decimal interchange code).***

8- bit code defining 256 symbols.
Still used in IBM mainframes and midrange systems.

ASCII

American National Standards Institute (ANSI) solution to representing symbols with bits of data was the ***ASCII*** character set.

Most commonly used in computers of all types.

Characters 0–31 and 127 are control characters. 32–64 special characters and numbers, 65-96 uppercase letters and a few symbols, 97-126 lowercase letters and a few common symbols.

Unicode

Evolving standard for data representation.
Uses 2 bytes (16-bits) to represent each letter, number or symbol.

2 bytes → can represent 65,536 different characters or symbols → enough for every unique character and symbol in the world (including vast Chinese, Korean and Japanese character sets as well as those found in known classical and historical texts).

Major advantage
Compatibility with ASCII codes. Unicode just extends the 256 character set of ASCII.

Developed in 1991 by Apple Computer Corporation and Xerox Corporation. It has been updated continually since. Autumn 1999 version 3 of the worldwide Unicode standard was released by the Unicode consortium. Version 3 includes a total of 57,709 16-bit code values.

Not yet universally adopted by software developers. If a single character set were available to cover all languages computer programs and data would be interchangeable.

How Computers process Data

2 components handle processing in a computer

1. CPU
2. Memory

Both are located on the motherboard, the circuit board that connects the CPU to the other hardware devices.

CPU

The CPU is the brain of the computer where data is manipulated. In the average microcomputer the entire CPU is a single unit called a microprocessor.

Most microprocessors are single chips mounted on a piece of plastic with metal wires attached to it. Some newer microprocessors include multiple chips and are encased in their own cover and fit into a special socket on the motherboard.

Every CPU has at least 2 basic parts

1. The control unit
2. The arithmetic logic unit

The Control Unit

All the computers resources are managed from the ***control unit***. It is the logical hub of the computer.

The CPU's instructions for carrying out commands are built into the control unit. The ***instruction set*** lists all the operations that a CPU can perform and is expressed in ***micro code*** (a series of basic directions telling the CPU how to execute more complex operations).

The Arithmetic Logic Unit

Computers data stored as numbers → much of the processing involves comparing numbers or carrying out mathematical operations.

The computer can perform 2 types of operations

1. ***Arithmetic operations***
2. ***Logical operations***

Arithmetic → Addition, Subtraction, Multiplication, Division
Logical → Comparisons such as Equality, Greater than, Less than

Every logical operator has an opposite.

Some of the logical operations can be done on text, e.g. searching for a word in a document means that the CPU carries out a rapid succession of *equals* operations to match the sequence of ASCII codes making up the search word.

Many instructions carried out by the control unit involve moving data. When the control unit encounters an instruction involving logic or arithmetic it passes this to the ***ALU (arithmetic logic unit)***

The ALU includes a group of ***registers*** (high speed memory locations built directly into the CPU). These are used to hold the data currently being processed.
E.g. Control unit loads 2 numbers from memory to registers in the ALU. Control unit tells ALU to divide the 2 numbers (arithmetic) or compare to see if they are equal (logic).

Machine Cycles

A CPU executes an instruction by taking a series of steps. The complete series of steps is called a ***machine cycle.***

Machine cycle can be broken down into

1. ***Instruction cycle***
2. ***Execution cycle.***

The instruction cycle has 2 steps

1. ***Fetching***

   Before the CPU executes an instruction the control unit must retrieve (***fetch***) a command or data from the memory.
2. ***Decoding***

   Before the command can be executed the control unit must break down (***decode***) the command into instructions corresponding to those in the instruction set.

The CPU is now ready to begin the execution cycle.

1. ***Executing***

   When the command is executed the CPU carries out the instructions in order by converting them to microcode.
2. ***Storing***

   CPU may be required to store the results of an instruction in memory (not always required).

The type of processor being used determines the number of steps in a machine cycle.

Although the process is complex the computer can accomplish it incredibly fast. CPU performance is often measured in ***millions of instructions per second (MIPS).***

Newer microprocessors perform faster by using a process called ***pipelining***. The control unit begins a new machine cycle – i.e. begins executing a new instruction – before the current cycle is completed. Executions are performed in stages, when the first instruction completes the fetching stage, it moves to the decode stage and a new instruction is fetched. Using this technique some microprocessors can execute up to 6 instructions simultaneously.

Memory

The CPU cannot store entire programs or large sets of data permanently. The CPU contains registers but these are small areas that hold only a few bytes at a time. The CPU needs to have millions of bytes randomly accessed space where it can quickly read or write programs and data while they are being used. This area is called ***memory***.

Physically consists of chips on the motherboard or on a small circuit board attached the the motherboard.

2 types of built in memory,
1. Permanent
2. Non permanent

Some memory chips always retain data they hold even when the computer is turned off → ***non-volatile***. Other chips, in fact most of the memory in a computer, lose their contents when the computer is turned off → ***volatile***.

ROM

Non-volatile chips always hold the same data which cannot be changed except through a special process that overwrites the data. Putting data into this type of memory is called **burning in the data** and is usually done in the factory.

During normal use the data in these chips is only read and used – not changed – so the memory is called ***read only memory (ROM)***.

One important reason for ROM – when the computer is switched on it needs to know what to do. ROM contains a set of start up instructions which ensures the rest of the memory is working correctly, checks for hardware devices and checks for an operating system.

RAM

Memory that can be instantly changed is called read-write memory or ***random access memory (RAM)***.

When people talk about memory in relation to a computer they usually mean the volatile RAM.

Purpose of RAM → hold data and programs while in use.

Physically consists of some chips on a small circuit board.

A computer does not need to search the entire memory each time it needs to find data because the CPU uses a ***memory address*** to store and retrieve each piece of data.

Memory address → number indicating a location on the memory chips. Start at 0 and go up to number of bytes – 1 in the memory.

Referred to as random access due to its ability to access each byte of data directly. ROM is random access as well so the name can be misleading. Tip for remembering → data in ROM does not change, data in RAM changes constantly.

Newer video cards, sound cards and printers have their own built in RAM.

2 types of RAM

1. ***Dynamic RAM (DRAM)***

   Gets it name from the fact that it must be refreshed frequently (refreshed → recharged with electricity). Must be refreshed many times per second or they will lose data.

2. ***Static RAM (SRAM)***

   Does not need to be refreshed as often and can hold data for longer.

SRAM is considerably faster than DRAM.
DRAM technologies typically support access times of approx 60ns. SRAM ships typically support access times of 10ns. As a consequence SRAM is more expensive than DRAM and not frequently used in PCs.

Flash Memory

Stores the data even when the computer is turned off. ROM is a form of flash memory. Other machines to use flash memory are digital cameras. You have to store the picture until transferred to your computer even when the camera is powered off.

Factors affecting processing speed

Circuitry design of a CPU determines its basic speed but several additional factors can make chips work faster.

Already introduced to registers and memory. Will now see how these and others – such as cache memory, clock speed and data bus – affect the speed.

How Registers affect speed

Registers in the first PCs could hold 2 bytes (16 bits) each. Most CPUs today have 32-bit registers. Newer PCs and high-end workstations have 64-bit registers.

Size of the registers (sometimes called ***word size***) indicates the amount of data the computer can work with at any given time. The bigger the word size the more quickly the computer can process the set of data. All other factors being kept equal, a CPU with 32-bit registers can process data twice as fast as one with 16-bit registers.

Memory and Computing Power

More RAM in a computer means that the computer can use bigger more powerful programs, which can access bigger data files.

More RAM means that the computer can run faster. The greater the amount of a program that the computer can load in to memory the faster it can run it. However a computer does not need to load an entire program into the memory to run it.

An example of this is Windows 98. A computer with 32MB of RAM can run Windows 98 even though the program occupies 195MB of disk storage space. The program loads the most essential parts into the memory and when it needs to access other parts of the program on the hard disk it can unload, or ***swap out*** nonessential parts from the RAM to the hard disk. It then can load or ***swap in*** the program code or data that it needs.

This method can result in a slow system performance. If your PC has 64MB (or more) of RAM you will notice a dramatic difference in how fast Windows 98 runs.

New PCs come with at least 64MB of RAM but 128MB is rapidly becoming the standard. Many new computers can hold as much as 1GB of RAM.

If you already own a PC you can increase the amount of RAM by plugging it into the motherboard. Chips are usually grouped together on small circuit boards called ***single in-line memory modules (SIMMS)*** or ***dual in-line memory modules (DIMMS)***. Each of these can hold from 1MB → 64MB and connect to the motherboard with 30-pin, 72-pin or 128-pin connections.

The Computer's Internal Clock

Every microcomputer has a ***system clock***, but it's main purpose is not to keep the time. It is driven by a quartz crystal. When electricity is applied the molecules vibrate millions of times per second at a rate that never changes. The speed of the vibrations is determined by the thickness of the crystal. The computer uses the vibrations to time its processing operations.

Over the years system clocks have become steadily faster. The first PC operated at 4.77MHz. (Hertz is a measure of cycles per second, MHz means *millions of cycles per second*)

The computers operating speed is tied to the speed of the system clock. A ***clock cycle*** is a single tick or the time it takes to turn a transistor off and back on. If a computers ***clock speed*** is 300MHz then it ***ticks*** 300 million times per second.

Clock speed has a tremendous impact on CPU performance. A CPU operating at 300MHz can process data almost twice as fast as one operating at 166MHz.

Clock speeds of 1GHz have already been achieved (early 2000).

The Bus

The term ***bus*** refers to the path between the components of a computer. There are 2 main buses in a computer

1. Internal (system) bus
2. External (expansion) bus

The system bus resides on the motherboard and connects the CPU to other devices residing on the motherboard. An expansion bus connects external devices such as keyboard, mouse, modem, printer etc. to the CPU. Cables from disk drives and other internal devices may also be plugged into a bus.

The system bus has 2 parts
1. Data bus
2. Address bus.

The Data Bus

An electronic path that connects the CPU, memory and other hardware devices on the motherboard.

The bus is actually a group of parallel wires. The number of wires affects the speed at which data can travel between hardware components. Each wire can transfer 1 bit at a time → 8-wire bus can transfer a full byte. 16-wire bus can transfer 2 bytes. Newer computers have a 64-bit data bus → transfer bytes at a time.

The Address Bus

A set of wires similar to the data bus which connects only the CPU and the RAM and carries only memory addresses. Today's CPUs have address buses wide enough to address 64GB of RAM.

Bus Types

Common Technologies

Industry Standard Architecture (ISA)
16-bit data bus. Became the industry standard on its release in the mid 80's and still used to attach slower devices (such as modems) to the CPU.

Local Bus
Developed to attach faster devices to the CPU. It is an internal system running between components on the motherboard. Most systems use some type of local bus couples to one or more types of expansion bus.

Peripheral Component Interconnect (PCI)
Type of local bus designed by Intel to make it easier to integrate new data types such as sound, video and graphics.

Accelerated Graphics Port (AGP)
Incorporates a special architecture allowing the video card to access the systems RAM directly → increases speed of graphic performance. Has led to the development of many types of accelerated video cards supporting 3D and full-motion video. Cannot be used with all PCs, the system must use a chip that supports the AGP standard. New computers feature AGP graphics capabilities as well as PCI system bus and an expansion bus.

Two new expansion bus technologies due to replace most existing buses in the future

1. ***Universal Serial Bus (USB)***
2. ***IEEE 1394*** (called ***Firewire*** on Macs).

Provide fast data transfer speeds and eliminate the need for expansion slots and boards. New PCs and Macs feature at least one USB port and each USB port can support 127 different devices.

In addition to the number of bits they can transfer at any one time buses are measured according to their ***data transfer rates*** – amount of data they can transfer in 1 second. Usually measured in ***megabits per second (Mbps)*** or ***megabytes per second (MBps)***.

E.g. USB data transfer rate of 12Mbps.
IEEE 1394 data transfer rate of 400Mbps.
AGP rates 266MBps but can support > 1GBps.
PCI rates 133MBps

Cache Memory

Moving data between RAM and CPUs registers is one of the most time consuming operations a CPU performs. The RAM is much slower than the CPU. Partial solution → include a cache memory. ***Cache memory*** is similar to RAM except it is extremely fast compared to normal memory and is used differently.

When a program is running and the CPU needs to read data or instructions from RAM it first checks to see if it is in the cache. If it is not there it reads the data from RAM into its registers and also loads a copy into the cache. Next time it needs this data it can read it from the cache so saving time.

Cache memory has been built into most PC CPUs since the late 80s. The CPU resident cache is often called ***Level-1 (L1) cache***. Today many CPUs have as much as 256kB of cache built in.

In addition cache is also added to the motherboard. This motherboard resident cache is often called ***Level-2 (L2) cache***. Many PCs come with 512 or 1024kB of L2 cache, higher-end systems can come with as much as 2MB of L2 cache.

Extending the Processor's Power to Other Devices

Need to know how to connect a new piece of hardware to the bus. In some cases you can plug the device into an existing socket or port on the back of the computer. Older computers feature only 3 or 4 distinct types of ports but newer systems provide a wide array of specialised ports. When a port is not available you need to install a circuit board that includes the port you want.

Serial and Parallel Ports

A ***parallel interface*** is a connection of 8 or more wires through which the data can flow simultaneously. Most computer buses transfer 32 bits simultaneously. However the standard parallel interface for external devices (such as printers) usually transfer 8 bits at a time over 8 separate wires.

A ***serial interface*** transmits data one bit at a time through a single wire, the interface includes additional wires to control the flow of data. Inside the computer a chip called a ***universal asynchronous receiver transmitter (UART)*** converts parallel data from the bus into serial data for the serial cable.

Parallel interfaces can handle a higher volume of data than a serial interface.

Other Ports

In addition to parallel and serial ports many computers include specialised ports

SCSI
Small Computer System Interface. Instead of forcing the user to plug multiple cards into the expansion slots a single SCSI adaptor extends the bus outside the computer by way of a cable. It is like an extension cord for the data bus. You can plug one SCSI device into another to form a chain.

Emerging standard → ***SCSI-3*** which can link as many as 127 devices. Ultra3 SCSI (newest standard) supports a 32-bit bus and transfer rates of 160Mbps.

To provide a SCSI port insert a SCSI adaptor board into one of the expansion slots.

Fast high quality hard disk drives often have SCSI interfaces as do scanners, tape drives and optical storage devices such as CD-ROM drives.

Plugging many devices into a single port is known as ***daisy chaining***.

USB
Mentioned before, is rapidly gaining popularity in PCs and Macs. Many experts believe the USB will emerge as the single bus standard of the future.

IEEE 1394
Also uses a single port. This technology is very expensive and so is not expected to become the dominant bus technology.

Musical Instrument Digital Interface (MIDI)
Has been in use since the early 80s when technology was created to enable electronic instruments to communicate. Has since been adapted to the PC and many sound cards are MIDI compliant. Using the MIDI port you can plug in a wide variety of musical instruments. MIDI systems are widely used in recording and performance to control settings for electronic synthesizers, drum machines etc.

Expansion Slots and Boards

PCs are designed so that users can configure the machines to their own particular needs. PC motherboards have 2 or more ***expansion slots*** – extensions to the computers bus – that provide a way of adding new components to the computer. The slots accept ***expansion boards*** (also called ***cards, adaptors*** or simply ***boards***).

Expansion slots on the motherboard are used for 3 main reasons

1. To give built in devices (such as hard disks and disk drives) access to the computers bus via controller cards.
2. To provide I/O ports on the back of the computer for external devices such as monitors, external modems, printers and game controllers.
3. To give special purpose devices access to the computer. E.g. An ***accelerator*** card – self contained device that enhances processing speed through access to the computers CPU and memory by way of the bus.

Adaptors that serve input or output purposes provide a port to which devices can be attached and act as translators between the bus and the device itself. Some adaptors also do a significant amount of data processing. E.g. A video card provides a port into which to plug the monitor but it also manages the video memory and handles the processing required to display images on the monitor. Other I/O devices commonly requiring a card in an expansion slot include sound cards, internal or fax modems, NIC (network interface cards) and scanners.

PC cards

Another type of expansion card (originally called a Personal Computer Memory Card International Association – or PCMCIA card) is a small device about the size of a credit card. Designed for notebooks and computers that are too small to accept standard expansion cards.

Fits into a slot at the back or side of the computer. These cards are used for a wide variety of reasons and can house modems, network cards, memory and hard disk drives.

3 categories of PC card technologies, Type I, Type II and Type III, each type typically defined by purpose.

Type I → usually contain memory.
Type II → usually network adaptors.
Type III → usually contain tiny hard drives.

Type I are the thinnest and have the fewest uses, Type III are the thickest and enable developers to fit disk storage devices into the card size shell.

Plug and Play

With the introduction of Windows 95 Intel based PCs began supporting ***plug and play*** standard. If you use hardware that complies with Window's plug and play standard then the operating system can detect new components automatically, check for existing drivers to run the new device and load the necessary files.

Windows will in some cases prompt to install needed files from a disk and may require a restart of the system for the new hardware setting to take place.